Weakly Alignment-free RGBT Salient Object Detection with Deep Correlation Network

Zhengzheng Tu, Zhun Li, Chenglong Li, Jin Tang

***Abstract*—RGBT Salient Object Detection (SOD) focuses on common salient regions of a pair of visible and thermal infrared images. Existing methods perform on the well-aligned RGBT image pairs, but the captured image pairs are always unaligned and aligning them requires much labor cost. To handle this problem, we propose a novel deep correlation network (DCNet), which explores the correlations across RGB and thermal modalities, for weakly alignment-free RGBT SOD. In particular, DCNet includes a modality alignment module based on the spatial affine transformation, the feature-wise affine transformation and the dynamic convolution to model the strong correlation of two modalities. Moreover, we propose a novel bi-directional decoder model, which combines the coarse-to-fine and fine-to-coarse processes for better feature enhancement. In particular, we design a modality correlation ConvLSTM by adding the first two components of modality alignment module and a global context reinforcement module into ConvLSTM, which is used to decode hierarchical features in both top-down and button-up manners. Extensive experiments on three public benchmark datasets show the remarkable performance of our method against state-of-the-art methods.**

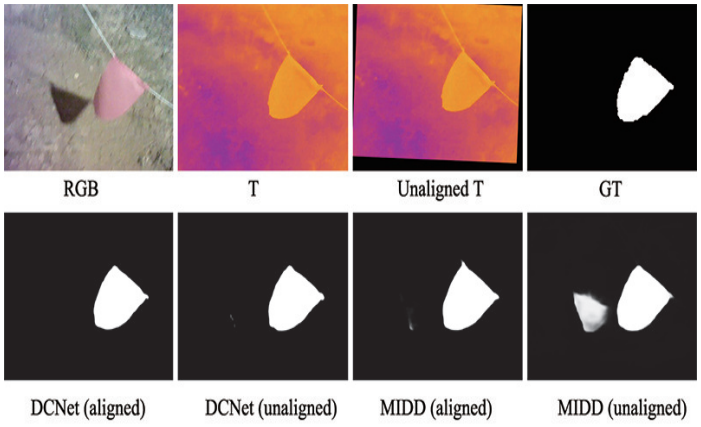


Fig. 1. Comparison of our method(DCNet) and MIDD [12] on aligned/unaligned RGBT image pair. With the original unaligned thermal image, DCNet shows robust performance while the MIDD [12] fails to counter the modalities bias.

I. Introduction

RGBT salient object detection focuses on segmenting the common salient objects or regions from paired visible light and thermal infrared images. Benefiting from the thermal sensors, the temperature field of objects contributes to the researches on visible light images, such as RGBT tracking [1], [2] and multi-spectral person Re-ID [3], [4]. For RGBT SOD, the foremost problem is to explore the correlation of two modalities, so as to find the common salient objects.

Existing researches mainly focus on fusing the modalities for information complementation. Some early works [5], [6], [7] propose traditional graph-based methods to infer saliency in RGBT images. Later, Tu et al. [8] propose to integrate multi-level deep features by a collaborative graph learning algorithm, which further improves the performance of RGBT SOD. However, these traditional methods depend on robust super-pixel segmentation results and cannot explore the correlation of modalities adequately. Several deep learning based methods [9], [10], [11], [12] have been proposed to fuse the multi-scale, multi-modal and multi-level features. They have made great progress on RGBT information complementation of modalities, and also achieve better performance.

It is worth mentioning that the visible-thermal cameras always capture the weakly aligned image pairs, caused by unaligned visible and thermal lenses. Meanwhile, manual image alignment is a labour-consuming work. However, all above-mentioned methods all rely on the well-aligned RGBT image pairs, and would thus have degraded performance when handling weakly aligned RGBT images. As shown in Fig. 1, MIDD [12] can not handle the unaligned scenarios well and achieves poor results.

In this work, we aim to address the task of weakly alignment-free RGBT SOD, and propose a novel deep correlation network, which builds the correspondences between modalities from spatial, feature and semantic levels, for weakly alignment-free RGBT SOD. Different from existing RGBT SOD methods [9], [10], [11], which focus on designing feature fusion models to integrate multi-modal or multi-level features, we concentrate on exploring multiple correlations between modalities and thus achieve more effective feature fusion even in unaligned image pairs. In specific, we design

This work was supported in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2020A0033, in part by Anhui Provincial Natural Science foundation under Grant 2108085MF211, in part by Anhui Energy Internet Joint Fund Project under Grant 2008085UD07, in part by the National Natural Science Foundation of China under Grant 61876002, in part by Anhui Provincial Key Research and Development Program under Grant 202104d07020008, and in part by the NSFC Key Project of International (Regional) Cooperation and Exchanges underGrant 61860206004. (*Corresponding author is Chenglong Li*)

Zhengzheng Tu, Zhun Li and Jin Tang are with Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: zhengzhengahu@163.com; lizhun118@foxmail.com; tangjin@ahu.edu.cn;senith@163.com)

Chenglong Li is with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230601, China, and also with the Institute of Physical Science and Information Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com)

a Modality Alignment Module (MAM) including the spatial affine transformation, the feature-wise affine transformation and the dynamic convolution operations. We adopt the spatial affine transformation to establish the correlation of object positions in two modalities. Then, we take the feature-wise affine transformation to establish the correlation of features. We further add a dynamic convolution component which performs on high-level features for semantic correlation. As the dynamic convolution operation needs a high memory cost, we just use a MAM followed by a Pyramid Pooling Module [13] to extract effective global contexts.

Moreover, most of existing SOD methods [14], [15], [16] adopt the progressive decoder, which integrates hierarchical features in a top-down manner. These methods perform a coarse-to-fine decoding process for perception and show effective performances. We consider that the fine-to-coarse perception process is also necessary for feature learning and can provide a novel perspective for feature decoding. Therefore, we combine these two perception processes and propose a bi-directional decoder to obtain a more powerful perception ability. In particular, based on the ConvLSTM [17], we propose a recurrent model that combines the modality correlation ConvLSTM with the original ConvLSTM to achieve an effective decoding process. In specific, we modify the ConvLSTM by adding the first two components of MAM and a global context reinforcement part into ConvLSTM, named as Modality Correlation ConvLSTM (MC-ConvLSTM). We use the MC-ConvLSTM as a decoding unit to conduct the coarse-to-fine and the fine-to-coarse perception. Finally, we integrate the outputs of two perception processes to refine the features by a ConvLSTM cell.

Benefited from the MAM and bi-directional decoder, our method has a strong ability to establish the correlation of two modalities and obtain the optimal feature representations.

Extensive experiments on three public available datasets prove our outstanding performance against the state-of-the-art RGBT SOD methods. The major contributions can be summarized as follows.

- We address **a new task** called weakly alignment-free RGBT SOD to relieve human labor and save time cost in RGBT SOD.
- We propose **a novel approach** based on the deep correlation network, which takes correlations between modalities into account from spatial, feature and semantic levels. It takes advantages of the information of two modalities as much as possible and achieves more effective representations.
- We design **an effective module** called Modality Alignment Module (MAM) to handle the problem of spatial misalignment of two modalities. By implicitly learning spatial affine transform and dynamically generating intermediate representation, MAM can robustly capture the spatial correlation of two modalities in challenging scenarios.
- We present **a novel bi-directional decoder** based on ConvLSTM and MC-ConvLSTM in a recurrent manner, which performs both coarse-to-fine and fine-to-coarse decoding processes, to enable our network have a great ability for information selection and suppression.

II. Related Work

A. RGBT Salient Object Detection

As mentioned above, the traditional methods [5], [6], [7] depend on robust super-pixel segmentation results and cannot present the correlation of modalities well.

Along with deep learning has shown the superior ability for feature representation, Tu et al. [9] build a large scale dataset for RGBT SOD, and they simultaneously provide an effective baseline method to integrate multi-level multi-modal features by a hierarchical attention mechanism. Zhang et al. [10] also design a network to fuse multi-modal information at various stages. Zhang et al. [11] revisit feature fusion for mining intrinsic patterns of RGBT saliency and design modules for the multi-scale, multi-modal and multi-level feature fusion. More recently, Tu et al. [12] present a multi-interactive dual decoder to implicitly restrain the bias regions by ground truth supervision and simultaneously infer the common salient regions. Zhou et.al [18] propose a cross-modality fusion module to fuse features of two modalities. Then they consider the salient object boundaries and design a bilateral reversal fusion module to fuse foreground and background information. And finally, they use a multilevel consistent fusion module to obtain complementary information by combining features across different levels. Later, Zhou et.al [19] further build a perceived importance fusion network with adversarial learning assistance. They use a progressively guided optimization structure to refine low-level features and design a perceived importance fusion module to weigh and fuse different modalities. Huo et.al [20] use a lightweight backbone and propose a context-guided cross modality fusion module to filter the noise and explore the complementation of two modalities. In addition, they also design a stacked refinement network to perform the interaction of semantic and spatial information. Wang et.al [21] propose a cross-scale alternate guiding fusion module to mine the high-level semantic information for global context, a guidance fusion module to achieve cross-modality fusion and a cross-guided fusion module for decoding in a cross-guided manner.

These methods have made great progress on RGBT multi-modal information complementation, which helps to achieve a better performance. However, all these above-mentioned methods are depend on well-aligned RGBT image pairs to learn the multi-modal information complementation. In this paper, we propose a novel deep network based on correlation of modalities for weakly alignment-free RGBT SOD.

B. Affine Transformation

The spatial affine transformation is a coordinate transformation, which maps the coordinate in source coordinate system to the target coordinate system by a transformation matrix. The spatial affine transformation applied to images is a rigid transformation that includes translation, rotation, shear-warp and even includes scaling operation. The spatial transformer network (STN) [22] realizes this process on the deep network

by predicting a transformation matrix. STN has been widely used in medical images registration [23], multi-modal images registration [24].

In our method, we use the spatial transformer network on features to obtain the fused feature with spatial correlation instead of input images for alignment. The reason is that we have no effective information for make explicit supervision and we are also unwilling to introduce extra information since we aim to design an universal method for weakly alignment-free RGBT SOD. Therefore, we embed the spatial transformer network to enable our method to learn spatial correlation in two modalities in an implicit manner.

The feature-wise affine transformation can be considered as a feature modulation mechanism. The parameters of affine transformation are generated from the features in one domain and then used to modulate the features in another domain. Therefore, the correlation between two domains can be learned during the training process. In a work of image super resolution [25], a spatial feature transform layer has been proposed for modulating features with semantic segmentation probability. They receive semantic segmentation probability to generate parameters of affine transformation for scaling and shifting feature maps of a specific layer. Chen et.al [26] design a progressive transformation network with semantic-aware style for blind face restoration. They embed the feature affine transformation into each restoration step, where they receive low quality image to generate affine transformation parameters to modulate a learned constant. In face deblurring task, Jung et.al [27] further propose a self spatial feature transform, in which they modulate features by the parameters generated from these features. Lang et.al [28] design a more novel mutual affine convolution that splits the input feature into several groups in channel dimension and modulates each group by those remaining feature except for the current feature. Finally, they concat these groups for obtaining modulated feature. This is also a self affine transformation.

In the salient object detection task, Tu et.al [16] use the edge prior to modulate encoded features by adopting the feature affine transformation. And Li et.al [29] use the feature affine transformation with gate mechanism embedded. They also use the edge prior to modulate encoded features. Inspired by these works, we use features of the visible light image to modulate the thermal image features, thus realizing the modality correlation on feature level.

C. Dynamic Convolution Operation

The earliest work of dynamic convolution is proposed for video and stereo prediction task [30], called dynamic filter. It designs the dynamic local filtering layer that generates sample specific and position specific kernels according to the inputs or feature maps. This work [30] has shown a powerful ability on establishing the semantic correlation between two features. Then, dynamic convolution is adopted to explore the correlation between text and image. Li et.al [31] use the dynamic convolution to generate target-specific visual filters depending on the text query, then use these filters to convolve the image feature for localizing the target in the video frames. Chen et.al [32] focus on referring image segmentation task where they also apply dynamic convolution to establish semantic correlation between language and visual domains. Dynamic convolution is not only used to learning semantic correlations across domains, but also widely applied to video frames for video interpolation [33], video denoising [34] and video deblurring [15]. These successful applications prove that dynamic convolution can effectively establish the semantic correlation between two images.

In recent RGBD SOD methods, Pang et.al [14] design the dynamic convolution layer to generate multi-scale filters for finding the correlation between two modalities. With the consideration that the kernel computation will introduce a large number of parameters and it is difficult to extend the kernel computation to multiple scales, Pang et.al [14] design a dynamic convolution operation with the characteristics of depth-wise separable convolution and dilated convolution. Chen et.al [35] consider that dynamic convolution can increase representation capability without increasing the network depth or width. They use multiple parallel convolution kernels and then dynamically aggregate them by attention mechanism. Dynamic convolution in this work [35] only introduces extra computational cost for computing attentions and aggregating kernels. More recently, Li et.al [36] think that property of dynamic convolution can not be limited to one dimension(kernel number) of the kernel space, should be expanded to other three dimensions including spatial size, input channel number and output channel number. Therefore, Li et.al [36] utilize a multi-dimensional attention mechanism to compute four types of attentions on four dimensions of the kernel space in a parallel manner.

With the dynamic speciality, dynamic convolution layer can flexibly handle different samples, indicating its great capacity on exploring the correlation between modalities. Inspired from the work[14], we think it is feasible to establish semantic correlation between visible features and thermal features by dynamic convolution operation. The variants proposed in [36], [35] aim at boosting the representation capability and can not meet our needs. And as discussed in [14],the original dynamic convolution operation needs a large memory cost. Limited by memory size of hardware, in our method, the dynamic convolution operation is only applied to the top encoded features for obtaining the reliable global context. Therefore, we adopt the dynamic convolution operator as another component of our MAM.

D. Decoding Strategy

The encoder-decoder framework is widely applied to pixel-wise segmentation or reconstruction task. The decoder can restore the resolution, simultaneously integrate useful information for the task. We consider that the decoder is a perception process actually, which includes useful information selection and useless information suppression. Most of existing methods [37], [38], [14], [15], [16] conduct the decoder progressively by aggregating encoded features in a top-down manner. This is a coarse-to-fine perception schema which can improve the details on the basis of general body region. In

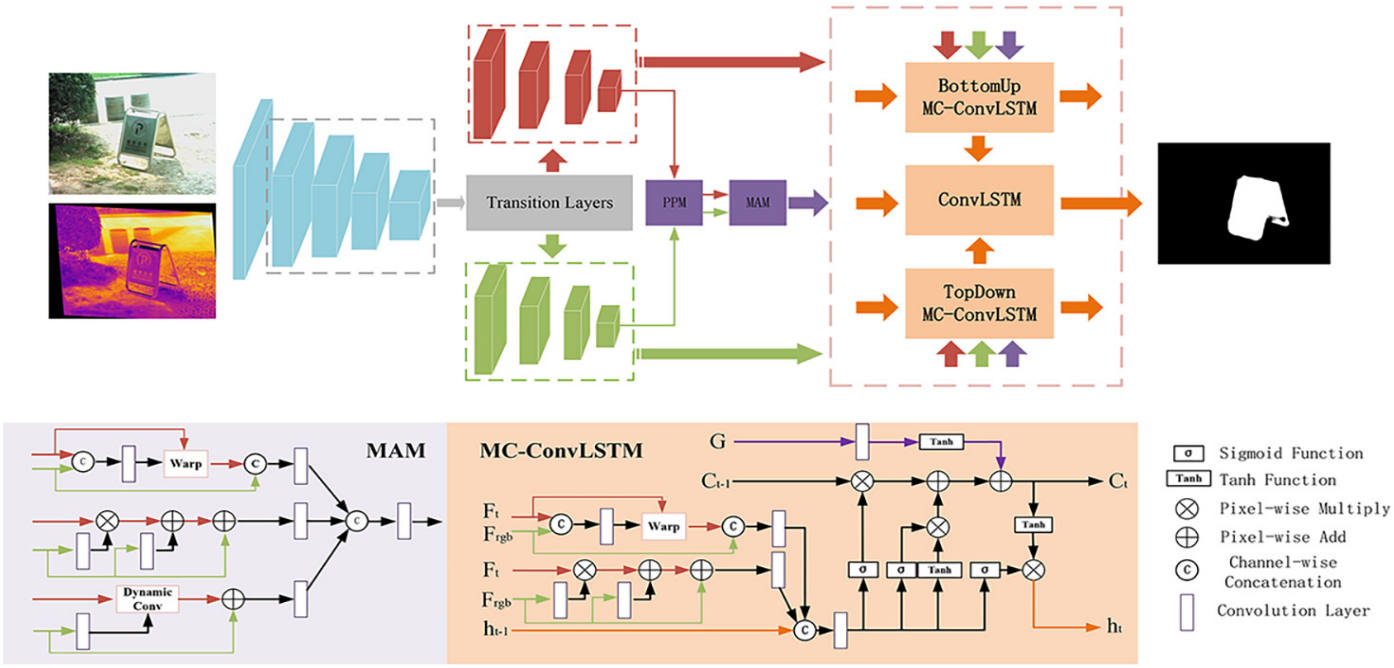


Fig. 2. The framework of DCNet can be divided into three parts. The first part is the encoder including a feature encoder and transition layers, it is used to encode hierarchical features of two modalities. The second part is about the global context, which contains a Pyramid Pooling Module(PPM) used in [12] and a designed Modality Alignment Module. The third part is the proposed bi-directional decoder with a ConvLSTM for refining features. And the final saliency map is predicted from the last refinement features.

our opinion, the fine-to-coarse perception is also a useful scheme that can infer and then fill the body region according to the existing details. Thus we propose a novel bi-directional decoder for performing information selection and suppression better.

III. Our Deep Correlation Network

A. Feature Encoder

In our method, we adopt a shared VGG16 [39] as the backbone encoder to extract hierarchical features from two modalities. To facilitate the subsequent operation, we further add a transition layer in each block in VGG16. The transition layer has a convolution layer with $3 * 3$ kernel size, 2 stride, 1 padding and 128 output channels, which halves the feature resolution and unifies the number of channels. Then it follows a batch normalization [40] and a Relu [41] for normalizing the data distribution and achieving non-linear activation.

Finally, through the above two procedures, we get four features with the same channel number in each modality. We mark these features of visible light image as $R_1 \sim R_4$ and thermal image as $T_1 \sim T_4$.

It should be noticed that our backbone model and the transition layers share parameters of two modalities. In our task, we have to consider the problem of spatial misalignment of two modalities, and thus adopt spatial affine transformation on encoded features in two modalities for capturing the fused features with spatial correlation. The encoded features of two modalities should be related with each other as much as possible. Therefore, we use shared encoder to extract common features.

B. Modality Alignment Module

As described before, the modality correlation is vital for RGBT SOD. When we establish the effective modality correlation, spatial deviation, modality bias and information complementation can be handled well. So we design a Modality Alignment Module (MAM) to establish stronger correlation of two modalities. The details of MAM can be seen in the bottom left of Fig. 2. In this section, we will introduce the detailed operations.

1) Spatial Affine Transformation Component: The spatial affine transformation focuses on establishing the correlation between spatial positions of two modalities, which enables the module to handle the weakly aligned data. In specific, we embed a spatial transformer network (STN) [22] into this component. The STN receives the corresponding features of two modalities, marked as F_r, F_t, and predicts a spatial affine matrix marked as M_p. M_p is a 2×3 matrix that defines the translation, rotation, shear-warp and scaling operations.

$$M_p = STN([F_r, F_t]) \tag{1}$$

where the STN represents our designed spatial transformation network, which stacks six convolution blocks with $8\times$ downsample, a global average pooling and a fully connected layer. The $[*]$ is the channel-wise concatenation. With the matrix M_p, we can easily get the warp filed φ by the integrated function of the Pytorch deep learning library [42]. Then we warp the F_t according to φ and obtain spatial transformed thermal feature marked as $\tilde{F}_t$.

$$\tilde{F}_t = \Phi(F_t, \varphi) \tag{2}$$

where $\Phi(*, *)$ is also the integrated warp function of Pytorch.

Finally, we concatenate the F_r and $\tilde{F}_t$ in channel-wise and fuse them by a convolution block with 64 output channels. The final output is a spatial correlated feature marked as F_s.

$$F_s = ConvBlock([F_r, \tilde{F}_t]) \tag{3}$$

where the $ConvBlock$ is the convolution block that contains a convolution layer, batch normalization and Relu activation function.

2) Feature-wise Affine Transformation Component: The feature-wise affine transformation component has a modulation mechanism that generates dynamic modulation parameters to establish feature-wise correlation between two features. This mechanism has a powerful ability to correlate the semantic information in two domains, which has been proven in [25], [16].

In this component, we also input the F_r and F_t to establish their correlation. With the same thinking as spatial affine transformation component, we use F_r to modulate F_t. We use two convolution blocks for F_r to generate the channel modulation parameters α and β respectively.

$$\alpha = ConvBlock(F_r) \tag{4}$$

$$\beta = ConvBlock(F_r) \tag{5}$$

Then we use α and β to transform F_t and obtain the thermal feature with the feature-wise affine transformation, marked as $\bar{F}_t$.

$$\bar{F}_t = \alpha * F_t + \beta \tag{6}$$

We directly add F_r and $\bar{F}_t$ with another convolution block to generate the final output feature, marked as F_f, that presents the feature-wise correlation.

$$F_f = ConvBlock(F_r + \bar{F}_t) \tag{7}$$

3) Dynamic Convolution Layer Component: The dynamic convolution layer is used to establish high-level semantic correlation between two modalities. We only adopt this component on the top encoded features because of the high memory cost of dynamic convolution operation and rich semantic information of high-level features.

As shown in Fig. 3, we use a convolution block on F_r to generate a dynamic filter $\kappa \in \mathbb{R}^{C*k^2 \times H \times W}$. At the position of each pixel, κ has $C * k^2$ channels, which can be reshaped to $\mathbb{R}^{C \times k \times k}$ that is used as a convolution kernel. Therefore, for thermal feature F_t, all the pixels have their own convolution kernel. Then we convolve F_t with generated kernels.

$$\kappa = ConvBlock(F_r) \tag{8}$$

$$\hat{F}_t = DynamicConv(F_t, \kappa) \tag{9}$$

where the $DynamicConv$ contains the filter reshape operation and the channel separation convolution. The $\hat{F}_t$ is the output feature. Finally, as same as feature-wise affine transformation component, we directly add F_r and $\hat{F}_t$ with another convolution block, so as to generate the final output feature, marked as F_d, that presents the semantic correlation between modalities.

$$F_d = ConvBlock(F_r + \hat{F}_t) \tag{10}$$

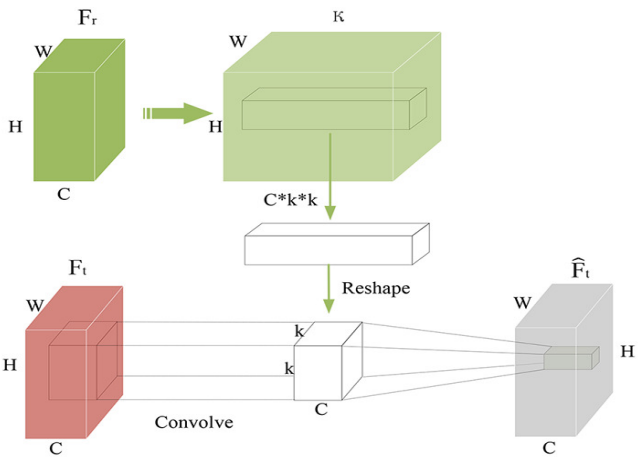


Fig. 3. The process of generating dynamic kernels and performing the dynamic convolution. H, W, C represent the height, width, channel number of input feature respectively and k is the kernel size of generated filter.

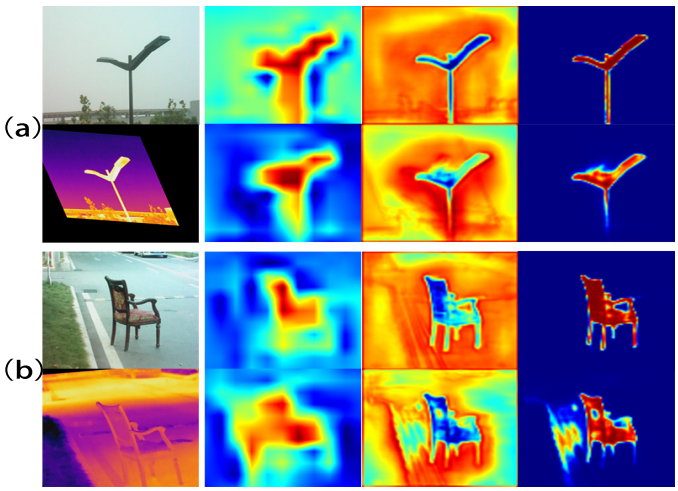


Fig. 4. We present two unaligned samples and visualize their global context scores, final refined features and saliency scores in the last three columns. In each example, the first row shows the results using spatial affine transformation component while the second row is the results without this component.

C. Global Context

The global context plays a significant role in SOD task, which highlights the spatial location of salient object, since it catches the semantic information as much as possible.

So we adopt our MAM followed by a Pyramid Pooling Module (PPM), which is used in [12] to obtain global context with stronger modality correlation. First, we respectively feed the top encoded features R_4 and T_4 into a shared PPM to obtain features G_r, G_t, with the global receptive filed.

$$G_r = PPM(R_4) \tag{11}$$

$$G_t = PPM(T_4) \tag{12}$$

Then, we feed G_r and G_t into our MAM. Through the embedded components, we get three features with different

modality correlations. We further concatenate these features in channel-wise and use a convolution block to fuse them.

$$G = MAM(G_r, G_t) \tag{13}$$

The final output G is the global context we need.

As shown in Fig. 4, without spatial position correlation of two modalities, the network learns to infer the salient region with strong dependence on visible light images since the visible light images are fixed with our setting. Therefore, the information of unaligned thermal image will be regarded as noises instead of useful cues. The network cannot use the information of corresponding thermal image effectively, thus obtains inaccurate results as shown in second row in each of two examples in Fig. 4. Along with the spatial position correlation established, the thermal information can also be used as a reference to infer more complete results.

D. Bi-directional Decoder

The common used decoder integrates hierarchical encoded features progressively in a top-down manner, which is a coarse-to-fine perception scheme. We consider that the fine-to-coarse perception scheme can provide a novel view for feature decoding, that is a process that fills the main regions according to the existing details. As ConvLSTM [17] is widely used for learning patterns of image sequences, we design a Modality Correlation ConvLSTM (MC-ConvLSTM) by adding the first two components in MAM and a global context reinforcement part into it. We use two MC-ConvLSTMs as perception units to learn two kinds of perception schemes. The module is shown in the bottom right of Fig. 2.

The computational processes of ConvLSTM is:

$$[w_i, w_f, w_o, w_g] = ConvBlock([In, h_{t-1}]) \tag{14}$$

$$i_t = \sigma(w_i), f_t = \sigma(w_f), o_t = \sigma(w_o), \tag{15}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(w_g) \tag{16}$$

$$h_t = o_t \circ \tanh(c_t) \tag{17}$$

Where i_t, f_t, o_t respectively refer to input, forget and output gates. C_t and h_t represent the cell state and the hidden state of current step. σ represents the sigmoid function. $\circ$ is pixel-wise multiply and In is input feature of current step.

In our method, we receive RGB image features(F_{rgb}), thermal features(F_t) and global context(G) every step. We embed the spatial transformation component and feature transformation component into ConvLSTM. F_s and F_f represent the output features of the two components. Then, In is the concatenation of F_s and F_f. And we add the global reinforcement part into ConvLSTM. Therefore, the C_t in MC-ConvLSTM becomes:

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(w_g) + \tanh(ConvBlock(G)) \tag{18}$$

In details, we use two shared MC-ConvLSTMs to respectively decode in a top-down and a bottom-up ways.

$$h_{Bt} = BottomUp(T_t, R_t, G, h_{Bt-1}), t = 1, 2, 3, 4 \tag{19}$$

$$h_{Tt} = TopDown(T_{5-t}, R_{5-t}, G, h_{Tt-1}), t = 1, 2, 3, 4 \tag{20}$$

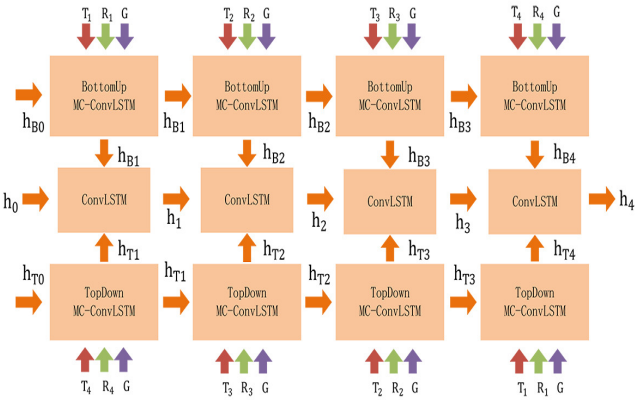


Fig. 5. The process of our bi-directional decoder.

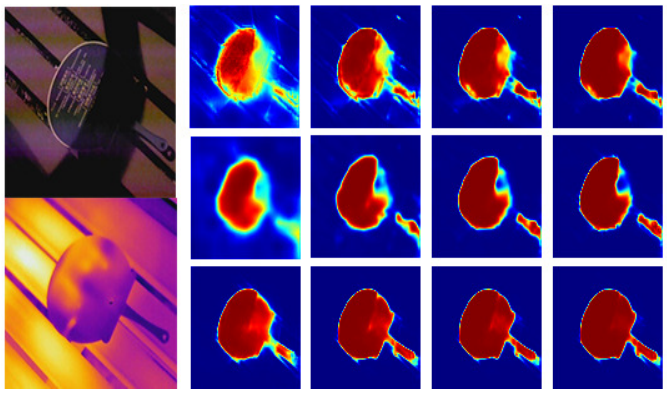


Fig. 6. The visualization of the decoding process. We respectively present the fine-to-coarse and coarse-to-fine perception schemes in the first and second rows. The third row is the results of our bi-directional decoder.

where $BottomUp(*,*,*,*)$ and $TopDown(*,*,*,*)$ represent two MC-ConvLSTMs respectively. The h_{Bt} and h_{Tt} are the hidden states of MC-ConvLSTMs, which are used as decoded features. We set all the initial hidden states and cell states to zero.

Then we use a normal ConvLSTM to receive two MC-ConvLSTMs' hidden states h_{Bt}, h_{Tt} for further refinement.

$$h_t = ConvLSTM(h_{Bt}, h_{Tt}, h_{t-1}), t = 1, 2, 3, 4 \tag{21}$$

Through the refinement $ConvLSTM(*,*,*)$, we get four decoded features $h_1 \sim h_4$. Finally, we adopt a score layer that is a $1 * 1$ convolution layer, there is one output channel on these decoded features. With the sigmoid function, we obtain four predicted saliency maps, marked as $S_1 \sim S_4$.

For the loss function, We simply adopt the binary cross entropy (BCE) loss on our four predicted saliency maps and a coarse saliency map predicted from global context.

We visualize the decoding process of a challenging sample in Fig. 6. The first row in Fig. 6 is a fine-to-coarse perception process. We can see that the main region is gradually filled by integrating the coarse information. But since the original encoded features with highest resolution contain too much noisy information, the false pixels are difficult to be corrected, leading to predict some false regions. The second row in Fig. 6 is a coarse-to-fine perception process. From this row, we can

find the details around the main region are gradually refined. But there is still a severe problem that the missing parts of salient region are less likely to be predicted if the original coarse main region is not complete. However, as shown in the last row, once we combine these two perception schemes, the results show noisy regions and missing parts become less. These phenomena prove that our view that the coarse-to-fine perception is a process that infers the details on the basis of general body region while the fine-to-coarse perception is a process that fills the body region according to the existing details. Combining these two kinds of perception schemes, the decoder has a greater potential to select information and suppress noise.

IV. Experiment

A. Experimental Setup

1) Dataset: We use the existing three RGBT SOD datasets to evaluate our method. VT821 [5],VT1000 [8] and VT5000 [9] collect 821,1000 and 5000 aligned RGBT image pairs respectively. We use the training set of VT5000 to train our network and compare the performances on VT821, VT1000 and VT5000's testing set. For proving our performance on unaligned image pairs, we randomly make the spatial affine transformation on all the testing data.

We use existing tools in Pytorch deep learning library [42] to perform this spatial affine transformation process. Similar to STN, we first define a transformation matrix $M_r \in R^{2\times 3}$.

$$M_r = \begin{bmatrix} s_x & r_x & t_x \\ r_y & s_y & t_y \end{bmatrix} \tag{22}$$

These six factors in matrix control the spatial transformation. Our task is aiming at weak alignment samples. For controlling the transformation, we set a value $\alpha > 1$ to adjust transformation intensity. We set random initial values for six factors with the range of $-1/\alpha \sim 1/\alpha$, which realizes the random transformations. s_x and s_y control the scaling process and can not be set as negative values. Therefore we add them with 1 to make their values around 1. In our experiment, we set α as a random value with range of $1 \sim 3$. Finally, we use Pytorch tools to generate warp field and to warp the thermal images. For the convenience of evaluation process, all the transformations are performed on thermal modality that is defined as the moved modality, meanwhile the visible modality is defined as fixed one.

2) Evaluation Metrics: We take the commonly used metric F-measure [43] to evaluate saliency maps. In our experiments, we binarize the predicted saliency map by a threshold that is diploid mean value, and then we compute F-measure and mark it as F_m. Moreover, we compute the weighted F-measure wF as same as the way in [44]. In addition, the mean absolute error(MAE) [45],S-measure(S_m) [46], E-measure(E_m) [47] are also used for the more comprehensive evaluation.

3) Implementation Details: We use Pytorch [42] to implement our method, which is trained on a single Titan Xp GPU. We adopt the stochastic gradient descent(SGD) [48] to optimize parameters with the weight decay of 5e-4, the momentum of 0.9. and the initial learning rate of 1e-3. Our network is trained 100 epochs with batch size of 4, the learning rate reduces to 1e-4 after 50th epoch. We resize the image to $352 * 352$ and randomly adopt horizontal flip for data augmentation. Then we directly divide the image by 255 to the range of 0 to 1. We further use the noisy data augmentation strategy [12] to enhance complexity of training samples. In order to conduct experiments on unaligned data, we randomly make spatial affine transformations on training set,by by randomly setting the value of the transformation matrix.

B. Comparison

1) Quantitative Evaluation: We compare our method with 6 advanced RGB SOD methods, including R3Net [49], PFA [50], CPD [51], EGNet [52], PoolNet [13] and BASNet [53]. We modify the input layer of these methods and concatenate two modalities as the input of 6 channels. Then we compare nine existing available RGBT SOD methods which contain three traditional graph-based methods that are SGDL[8], MTMR[5], M3S-NIR[7] and six deep learning based RGBT SOD methods that are ADF [9],MIDD [12],APNet [19],ECFFNet [18], CSRNet [20] and CGFNet [21]. In Table I, we make a quantitative comparison on five above-mentioned metrics. We can see that our method on aligned test datasets shows the remarkable performance against the state-of-the-art RGB/RGBT SOD methods. Compared with most recent deep learning based methods, our method have similar or superior performance. And the scores on five evaluation metrics are maintained in top three on the whole.

In Table II, we further train our method on training set with random spatial affine transformation augmentation and test it on unaligned data which is generated from three public RGB-T SOD datasets with random spatial affine transformation. We select seven compared methods including R3Net, EGNet, SGDL, ADF, MIDD, CGFNet and ICNet [54]. ICNet is a co-saliency method which focus on seeking same object in a group of images. It also needs to build semantic correlations of images. In our experiment, we apply it to RGB-T image pairs and use it to seek salient object in visible light image that is consistent with our experimental purpose. As shown in Table II, though our method aims at the unaligned image pairs, our method still performs best compared with those methods on unaligned data, with five metrics. And even comparing with SOTA methods on aligned data in Table I, our performance is also competitive. All of R3Net, EGNet, SGDL and ADF simply fuse features of two modalities and don't consider the correlation between them. Therefore their performances on unaligned data are relatively poor. As we can see in Table I, CGFNet has similar performance with our proposed method. However, in Table II, it has obvious gap compared with our method on unaligned datasets, which proves that the ability of our method to deal with spatial misalignment.

From Table I and Table II,We can observe the rate of performance decline of all methods from being tested on aligned data to being tested on unaligned data. And we compute the average decline rate on three datasets to

TABLE I
PERFORMANCE COMPARISON WITH 12 METHODS ON THREE TESTING DATASETS. THE BEST SCORES ARE HIGHLIGHTED IN **RED**, THE SECOND BEST SCORES ARE HIGHLIGHTED IN **GREEN**, AND THE THIRD BEST SCORES ARE HIGHLIGHTED IN **BLUE**.

Methods	VT821					VT1000					VT5000-Test				
	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF
PFA	0.756	0.761	0.592	0.096	0.526	0.809	0.813	0.688	0.078	0.635	0.737	0.748	0.563	0.099	0.498
R3Net	0.803	0.782	0.681	0.081	0.656	0.903	0.886	0.835	0.037	0.831	0.856	0.812	0.729	0.059	0.703
BASNet	0.856	0.823	0.735	0.067	0.716	0.923	0.909	0.847	0.030	0.861	0.878	0.839	0.764	0.054	0.742
PoolNet	0.811	0.788	0.652	0.082	0.573	0.852	0.849	0.751	0.063	0.690	0.809	0.788	0.643	0.080	0.570
CPD	0.843	0.818	0.718	0.079	0.686	0.923	0.907	0.863	0.031	0.844	0.894	0.855	0.787	0.046	0.748
EGNet	0.856	0.830	0.726	0.063	0.662	0.922	0.910	0.848	0.033	0.817	0.888	0.853	0.775	0.050	0.712
MTMR	0.815	0.725	0.662	0.108	0.462	0.836	0.706	0.715	0.119	0.485	0.795	0.680	0.595	0.114	0.397
M3S-NIR	0.859	0.723	0.734	0.140	0.407	0.827	0.726	0.717	0.145	0.463	0.780	0.652	0.575	0.168	0.327
SGDL	0.847	0.765	0.730	0.085	0.583	0.856	0.787	0.764	0.090	0.652	0.824	0.750	0.672	0.089	0.559
ADF	0.842	0.810	0.716	0.077	0.627	0.921	0.910	0.847	0.034	0.804	0.891	0.864	0.778	0.048	0.722
MIDD	0.895	0.871	0.804	0.045	0.760	0.933	0.915	0.882	0.027	0.856	0.897	0.868	0.801	0.043	0.763
APNet	0.907	0.867	0.816	0.034	0.792	0.938	0.921	0.883	0.021	0.883	0.914	0.876	0.820	0.035	0.807
ECFFNet	0.902	0.877	0.810	0.034	0.801	0.930	0.923	0.876	0.021	0.885	0.906	0.874	0.807	0.038	0.802
CSRNet	0.908	0.885	0.830	0.038	0.821	0.925	0.918	0.877	0.024	0.878	0.905	0.868	0.811	0.042	0.797
CGFNet	0.912	0.881	0.845	0.038	0.829	0.944	0.923	0.906	0.023	0.900	0.922	0.883	0.852	0.035	0.831
DCNet	0.913	0.877	0.841	0.033	0.822	0.949	0.923	0.911	0.021	0.902	0.921	0.872	0.847	0.035	0.819

TABLE II
PERFORMANCE COMPARISON WITH 6 METHODS ON THREE TESTING DATASETS WITH RANDOM SPATIAL AFFINE TRANSFORMATIONS. THE BEST SCORES ARE HIGHLIGHTED IN **RED**, THE SECOND BEST SCORES ARE HIGHLIGHTED IN **GREEN**, AND THE THIRD BEST SCORES ARE HIGHLIGHTED IN **BLUE**.

Methods	Unaligned-VT821					Unaligned-VT1000					Unaligned-VT5000-Test				
	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF	Em	Sm	Fm	MAE	wF
R3Net	0.760	0.727	0.596	0.099	0.565	0.841	0.815	0.729	0.059	0.710	0.803	0.729	0.603	0.078	0.565
EGNet	0.794	0.764	0.632	0.094	0.561	0.859	0.835	0.741	0.061	0.696	0.832	0.774	0.664	0.076	0.583
ICNet	0.859	0.831	0.739	0.047	0.727	0.911	0.888	0.836	0.037	0.834	0.846	0.793	0.703	0.064	0.674
SGDL	0.807	0.728	0.658	0.098	0.502	0.841	0.759	0.729	0.096	0.592	0.786	0.711	0.603	0.102	0.476
ADF	0.727	0.709	0.566	0.157	0.475	0.826	0.827	0.713	0.088	0.665	0.816	0.793	0.669	0.088	0.593
MIDD	0.873	0.840	0.756	0.059	0.707	0.914	0.896	0.841	0.034	0.814	0.878	0.844	0.763	0.052	0.719
CGFNet	0.876	0.856	0.783	0.066	0.739	0.921	0.911	0.862	0.031	0.829	0.892	0.860	0.794	0.050	0.737
DCNet	0.908	0.860	0.826	0.036	0.799	0.943	0.915	0.902	0.023	0.889	0.908	0.854	0.824	0.041	0.790

show the ability of each method to deal with unaligned data. On five metrics(Em,Sm,Fm,MAE and wF), EGNet declines $6.8\%, 8.5\%, 13.3\%, 37.7\%$ and 16.1%, MIDD declines $2.2\%, 2.8\%, 5.1\%, 20.5\%$ and 5.9%. CGFNet averagely declines $3.2\%, 2.3\%, 6.8\%, 50.5\%$ and 11.2%, which shows its instability when encountering the samples with spatial misalignment, also reflects its superiority with S-measure metric. Our method only declines $0.8\%, 1.6\%, 1.8\%, 10.5\%$ and 2.6%, which shows our method can retain the performance better when it encounters the unaligned image pairs.

C. Challenge-based Comparison

VT5000 has 11 challenges, including: 1) BSO(big salient object): the number of pixels in the salient object to total pixels is more than 26%; 2) SSO(small salient object):the number of pixels in the salient object to total pixels is less than 5%; 3) MSO(multiple salient object):the number of salient objects is more than 1; 4) LI(low illumination):the samples are collected in low illumination environment; 5) CB(center bias):the salient object is far away from the center of the image; 6) CIB(cross image boundary):the salient object crosses the boundaries of image; 7) SA(similar appearance):the salient object is similar to the background; 8) TC(thermal crossover):the salient object has similar temperature to other objects or its surrounding; 9) IC(image clutter): the background is cluttered; 10) OF(out of focus):the image is out-of-focus and blurred; 11) BW(bad weather):the image is collected in rainy or foggy days. In addition, VT5000 annotates the quality of two modalities. As shown in Table. III, we make a detailed comparison of fifteen methods on eleven challenges and two modality quality annotations provided by VT5000 testing set. We compute the mean $Fscore$ as the metric index which can mainly present the performance on challenges. We compare our method with six RGB SOD methods and eight RGBT SOD methods including three traditional graph-based methods and five deep learning based methods.

TABLE III
PERFORMANCE COMPARISON OF 12 METHODS ON 11 CHALLENGES AND 2 MODALITY QUALITY CATEGORIES. THE BEST SCORES ARE HIGHLIGHTED IN **RED**, THE SECOND BEST SCORES ARE HIGHLIGHTED IN **GREEN**, AND THE THIRD BEST SCORES ARE HIGHLIGHTED IN **BLUE**.

	BSO	CB	CIB	IC	LI	MSO	OF	SSO	SA	TC	BW	Bad RGB	Bad T
PFA	0.735	0.616	0.670	0.615	0.656	0.617	0.647	0.376	0.606	0.625	0.543	0.636	0.618
R3Net	0.808	0.763	0.799	0.714	0.764	0.754	0.735	0.623	0.701	0.701	0.722	0.713	0.692
BASNet	0.838	0.781	0.803	0.750	0.811	0.774	0.792	0.679	0.739	0.764	0.738	0.766	0.760
PoolNet	0.757	0.662	0.695	0.661	0.703	0.649	0.706	0.543	0.672	0.660	0.670	0.683	0.658
CPD	0.852	0.811	0.831	0.781	0.823	0.804	0.802	0.722	0.804	0.787	0.765	0.781	0.777
EGNet	0.849	0.807	0.832	0.784	0.822	0.786	0.791	0.658	0.761	0.763	0.743	0.766	0.749
MTMR	0.490	0.471	0.421	0.450	0.547	0.495	0.573	0.634	0.539	0.462	0.491	0.533	0.452
M3S-NIR	0.499	0.463	0.450	0.442	0.555	0.469	0.569	0.508	0.505	0.441	0.511	0.508	0.430
SGDL	0.677	0.658	0.622	0.625	0.659	0.660	0.688	0.711	0.594	0.620	0.588	0.601	0.606
ADF	0.853	0.815	0.835	0.790	0.833	0.805	0.806	0.742	0.799	0.796	0.774	0.788	0.789
MIDD	0.873	0.837	0.852	0.797	0.852	0.816	0.828	0.740	0.815	0.810	0.770	0.815	0.797
APNet	0.883	0.843	0.862	0.814	0.871	0.825	0.845	0.753	0.846	0.821	0.809	0.835	0.806
ECFFNet	0.876	0.831	0.868	0.803	0.848	0.823	0.821	0.723	0.809	0.808	0.756	0.810	0.798
CSRNet	0.868	0.812	0.834	0.791	0.855	0.815	0.837	0.759	0.799	0.801	0.760	0.824	0.792
DCNet	0.887	0.853	0.865	0.821	0.869	0.835	0.850	0.767	0.846	0.839	0.816	0.833	0.834

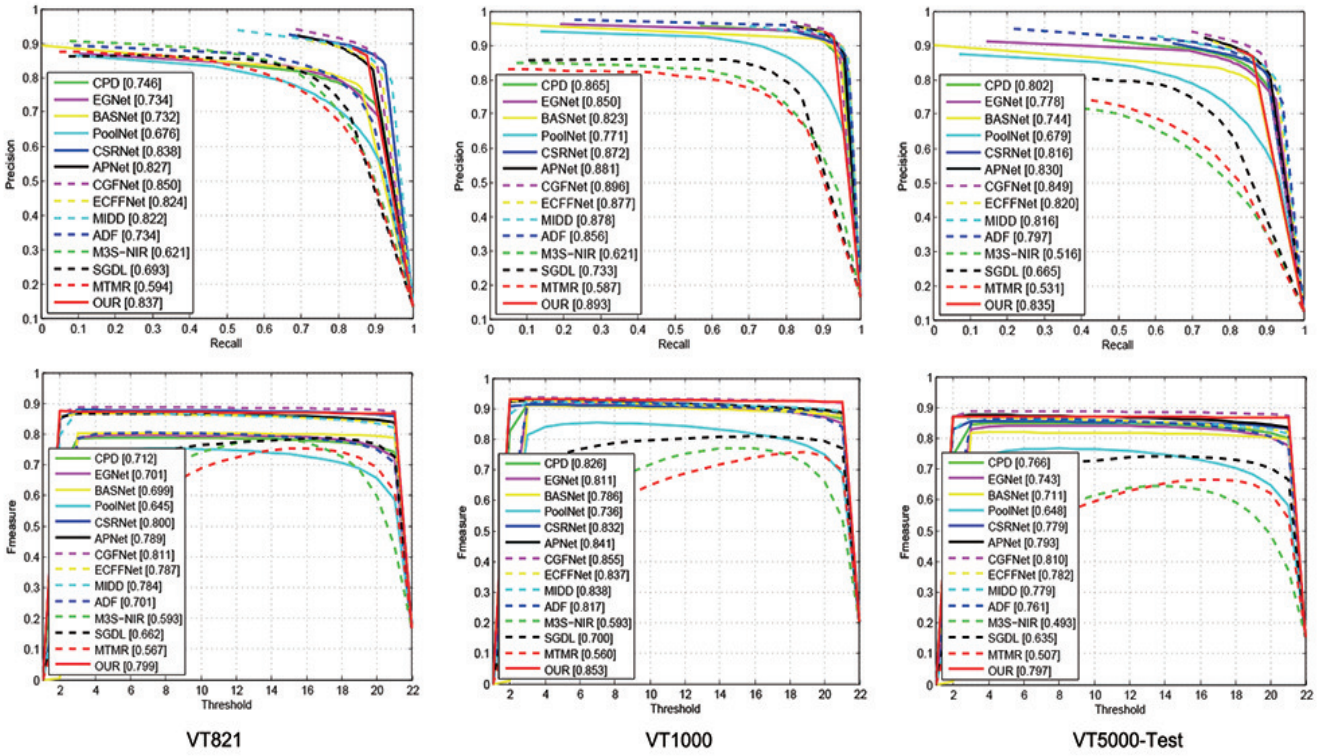


Fig. 7. The PR curve and F-measure curve of our results and 11 compared methods. The first raw is PR curve and the second raw is F-measure curve.

Over all challenges, the deep learning based RGBT methods show better performances than traditional methods and RGB methods, which proves that the importance of feature representation and modality correlation. The performances of our method on all challenges surpass ADF, MIDD and CSRNet. Compared with APNet and ECFFNet, our method shows better performance on samples with low illumination, similar appearance and defective visible image, and APNet and ECFFNet are more robust for dealing with incomplete salient objects. From a overall view, our method can keep optimal performance on all the challenges. It should be noticed that we do not specially design this network to handle any challenges but to make stronger modality correlation and decoding ability. Therefore, it is obvious that our remarkable performance is only relied on the better feature representation and information utilization, which exactly proves the effectiveness of our innovations.

However, we find that the IC,SSO and BW are the top three difficult challenges since all the methods show the lower scores on them. For dealing with these three challenges, the best way is to specially design suitable modules, which should be further studied in the future.

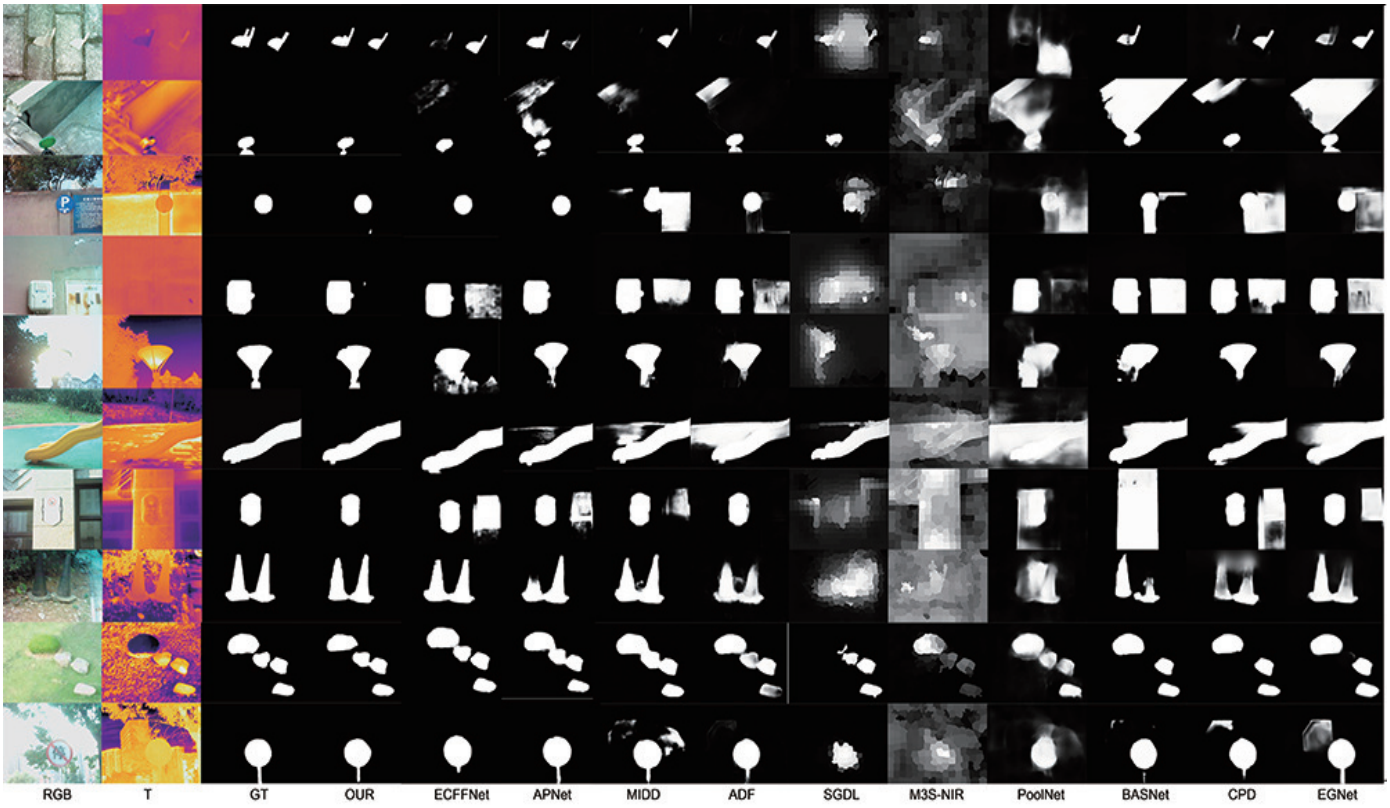


Fig. 8. The comparison of saliency maps on 10 selected challenging samples. We compare six RGBT SOD methods and three RGB SOD methods

TABLE IV
ABLATION EXPERIMENTS OF THREE MAIN COMPONENTS IN OUR DCNET ON VT821 AND UNALIGNED VT821. THE SAT,FAT AND DCL RESPECTIVELY REPRESENT SPATIAL AFFINE TRANSFORMATION, FEATURE-WISE AFFINE TRANSFORMATION AND DYNAMIC CONVOLUTION LAYER. THE OPTIMAL RESULTS ARE HIGHLIGHTED IN **BOLD**.

			VT821					Ualigned-VT821				
SAT	FAT	DCL	Fm	wF	Em	Sm	MAE	Fm	wF	Em	Sm	MAE
			0.807	0.789	0.890	0.857	0.045	0.790	0.734	0.877	0.815	0.047
✓			0.823	0.803	0.902	0.865	0.040	0.815	0.782	0.898	0.851	0.041
	✓		0.825	0.806	0.901	0.868	0.040	0.812	0.773	0.896	0.842	0.042
		✓	0.815	0.799	0.898	0.862	0.042	0.797	0.755	0.882	0.832	0.044
✓	✓		0.828	0.811	0.903	0.870	0.038	0.816	0.787	0.904	0.853	0.040
✓	✓	✓	**0.841**	**0.822**	**0.913**	**0.877**	**0.033**	**0.826**	**0.799**	**0.908**	**0.860**	**0.036**

D. Qualitative Evaluation

We select 10 challenging samples and illustrate their saliency maps in Figure. 8. Compared with six RGBT SOD methods and three advanced RGB SOD methods, our method shows more accurate segmentation of salient region. These three RGB SOD methods can predict clear results but cannot correlate two modalities, thus lead to segmentation with false region or missing region(1st,2nd,3rd,4th and 9th rows in Figure. 8). Two traditional graph-based methods cannot handle the complex cases such as low contrast(5th and 7th rows in Figure. 8) and cluttered background(6th and 10th rows). The four compared deep learning based RGBT SOD methods show better performances than the above two kinds of methods. As they consider the fusion of modalities, they can handle the influence of bad modality quality(4th,7th and 10th rows) in some extent. But they cannot establish the stronger correlation between modalities. Therefore, they may make some mistakes on samples with modality bias(1st,2nd,4th and 6th rows). On these challenging samples, benefited from our MAM and bi-directional decoder, our results can accurately focus on common salient regions.

E. Analysis for PR curves and F-measure curves

The PR curves and F-measure curves of our results and 13 compared methods are presented in Figure. 7. As shown in first row, our method has less variance on both $Precision$ and $Recall$ than compared methods. Regardless of the scale of the threshold, our method gets generally consistent results with metric indexes, in the range of $0.85 \sim 0.9$. This result demonstrates that our saliency maps have higher confidence

TABLE V
THE COMPARISON OF OUR METHOD AND EXISTING METHODS ON GFLOPS AND PARAMTERS. THE OPTIMAL RESULTS ARE HIGHLIGHTED IN **BOLD**.

	PoolNet	EGNet	ADF	MIDD	CGFNet	DCNet
GFLOPs	299.16	726.28	247.30	216.74	347.43	**207.31**
Paramters(MB)	200.34	426.38	317.15	200.09	266.93	**91.88**

in foreground regions. Though the threshold is close to 1, our results still have stable performances while the other methods decline a lot.

In the second row of Figure. 7, it is obvious that our F-measure scores surpass all the compared methods. The slope of our curve is nearly zero which is also proven that our results have higher confidence. This is benefited from our bi-directional decoder that learns two perception schemas, thus leading the higher confidence than the other methods with the general decoder. We will further prove this conclusion in subsequent ablation experiments.

F. Analysis for Network Complexity

Our proposed method uses an bi-directional decoder and a refined stream to perform decoding process in a recurrent manner, which has not been explored yet. It is necessary to test the computational complexity of our method to verify the feasibility. We compute the GFLOPs and network size of our method and some advanced methods for comparison. The GFLOPs is 10^9 FLOPs(floating point operations) that can be used for measuring the computational quantity of each sample. The network size is to measure the number of parameters of the network and its unit is MB(mega byte). As shown in Table V, our DCNet has lower GFLOPs and smaller network size, which proves that our novel decoder strategy is feasible and its computational cost is acceptable. As using the bi-directional decoder and a refined stream, our method has no great superiority on computational complexity. But it has obvious superiority on network size and it declines 54.08% compared with the suboptimal method.

G. Ablation Study

We respectively verify the effectiveness of components in our method, that are modality alignment module and bi-directional decoder. Moreover, we further make an ablation study on the proposed three modality correlation components.

1) Effectiveness of the modality alignment module: In Table VI, the method "w/o MAM" uses a convolution block instead of MAM to integrate the channel-wise concatenate G_r and G_t. Compared with the last row of Table VI, it shows $1.6\%, 1.7\%, 1.6\%, 1.0\%$ and 2.1% reductions on five metric indexes. Our network uses the MAM to obtain accurate global context that plays an important role in salient region location. Although the global features have the lower resolution, so that it is difficult to refine them, our proposed MAM can further improve the global context and obtain better performance.

TABLE VI
ABLATION EXPERIMENT FOR OUR MODALITY ALIGNMENT MODULE(MAM) AND BI-DIRECTIONAL DECODER(BD) ON VT821. THE OPTIMAL RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	VT821				
	Fm	wF	Em	Sm	MAE
w/o Mc-ConvLSTM	0.782	0.715	0.834	0.801	0.051
w/o BD	0.771	0.740	0.880	0.826	0.047
w/o TopDown	0.825	0.804	0.906	0.865	0.038
w/o BottomUp	0.822	0.802	0.903	0.865	0.039
w/o MAM	0.828	0.808	0.899	0.868	0.040
DCNet	**0.841**	**0.822**	**0.913**	**0.877**	**0.033**

2) The effect of spatial affine transformation: Comparing the first row and second row in Table IV, with the spatial affine transformation component embedded into MAM and Mc-ConvLSTM, the performances on aligned and unaligned VT821 dataset are obviously improved. Moreover, we can see that the gains on unaligned data($1.9\%, 1.7\%, 1.3\%, 0.9\%$ and 12.5%) are more than on aligned data($2.7\%, 6.1\%, 2.3\%, 4.2\%$ and 14.6%), which prove that the ability of spatial affine transformation component for establishing the spatial position correlation between two modalities.

Without the spatial affine transformation component, the network learns to infer the salient regions with strong dependence on visible light image since the visible light images are fixed images in the setting of our method. Once the spatial position correlation established, the thermal information can also be used as a reference to infer more complete results.

3) The effect of feature-wise affine transformation: Comparing the first row and the third row in Table. IV, the performances on two kinds of datasets are both improved. The gains on two datasets are $2.1\%, 2.1\%, 1.2\%, 1.3\%, 12.5\%$ and $2.7\%, 5.0\%, 2.1\%, 3.2\%, 11.9\%$ respectively on five metric indexes, which proves that the feature-wise affine transformation can essentially improve the ability of feature representation. In some difficult cases, one of the modalities catches little useful information. The spatial affine transformation component cannot establish spatial position correlation since lacking of reliable information. The feature-wise affine transformation component can establish feature-wise correlation which relieves the error introduced by spatial position correlation. With the collaboration of these two components, our network can tackle most of difficulties.

4) The effect of dynamic convolution layer: Comparing the first and the 4th row in Table. IV, the results show that the dynamic convolution layer can improve the performance. Comparing the 5th and the last row, though the other two components are embedded into MAM, the performance can still be further improved. These two comparisons determine that the dynamic convolution layer can establish another kind of correlation between modalities, which is different from the first two components. Limited from the memory cost, we only use this component to obtain global context. Nevertheless, it also effectively improves the predicted results.

5) Effectiveness of the bi-directional decoder: We make an exhaustive ablation experiments for our bi-directional decoder.

Firstly, we replace our Mc-ConvLSTM by the ConvLSTM to verify the effectiveness of our kernel perception unit("w/o Mc-ConvLSTM" in Table. VI). Compared with our DCNet, it has $7.5\%, 15.1\%, 9.5\%, 9.5\%$ and 35.3% reductions on five metric indexes respectively, because without the Mc-ConvLSTM, the modality correlation is weaken and the global context is scarce.

We further use the ConvLSTM for decoding in top-down manner together with using global context, marked as "w/o BD" in Table. VI. We can find that the performance also declines a lot without the bi-directional decoder. The "w/o TopDown" and "w/o BottomUp" respectively represent that we use a fine-to-coarse schema and a coarse-to-fine schema. As shown in Table. VI, when we solely use one of them, we get close performances. But when we combine these two perception schemas, the performance increases $2.1\%, 2.3\%, 1.0\%, 1.4\%$ and 2.1% on five metric indexes respectively. This observation further proves that our bi-directional decoder has greater potential to select and suppress information.

V. Conclusion

In this paper, we propose a novel deep correlation network for weakly alignment-free RGBT salient object detection. For the consideration of fact that the RGBT image pairs are spatially unaligned, we propose to establish multiple modality correlations between two modalities. Therefore not only the information complementation is achieved, but also the modality bias is suppressed. Then, we present a novel view on decoder that we propose a bi-directional decoder combined coarse-to-fine and fine-to-coarse perception schemes, which provides powerful potential for information selection and suppression. Our experiments show the effectiveness of the proposed methods and superior performance against state-of-the-art methods.

References

[1] C. Li, A. Lu, H. A. Zheng, Z. Tu, and J. Tang, "Multi-adapter rgbt tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[2] C. Li, L. Liu, A. Lu, Q. Ji, and J. Tang, "Challenge-aware rgbt tracking," in *European Conference on Computer Vision*, 2020.

[3] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018.

[4] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: Hypersphere manifold embedding for visible thermal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[5] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Chinese Conference on Image and Graphics Technologies*, 2018.

[6] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "Rgbt salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2019.

[7] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3s-nir: Multi-modal multi-scale noise-insensitive ranking for rgb-t saliency detection," in *IEEE Conference on Multimedia Information Processing and Retrieval*, 2019.

[8] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2019.

[9] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *arXiv preprint arXiv:2007.03262*, 2020.

[10] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2020.

[11] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[12] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for rgb-thermal salient object detection," *IEEE Transactions on Image Processing*, 2021.

[13] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *European Conference on Computer Vision*, 2020.

[15] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[16] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[17] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, pp. 802–810, 2015.

[18] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[19] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Apnet: Adversarial learning assistance and perceived importance fusion network for all-day rgb-t salient object detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[20] F. Huo, X. Zhu, L. Zhang, Q. Liu, and Y. Shu, "Efficient context-guided stacked refinement network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[21] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "Cgfnet: Cross-guided fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[22] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.

[23] M. C. Lee, O. Oktay, A. Schuh, M. Schaap, and B. Glocker, "Image-and-spatial transformer networks for structure-guided image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 337–345.

[24] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[25] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 606–615.

[26] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K.-Y. K. Wong, "Progressive semantic-aware style transformation for blind face restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 896–11 905.

[27] S. H. Jung, T. B. Lee, and Y. S. Heo, "Deep feature prior guided face deblurring," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3531–3540.

[28] J. Liang, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Mutual affine network for spatially variant kernel estimation in blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4096–4105.

[29] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3069–3082, 2020.

[30] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, pp. 667–675, 2016.

[31] Z. Li, R. Tao, E. Gavves, C. G. Snoek, and A. W. Smeulders, "Tracking by natural language specification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6495–6503.

[32] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, and M.-H. Yang, "Referring expression object segmentation with caption-aware consistency," in *British Machine Vision Conference (BMVC)*, 2019.

[33] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 670–679.

[34] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2502–2510.

[35] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 030–11 039.

[36] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," in *International Conference on Learning Representations*, 2021.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[43] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[44] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[45] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[46] D. P. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[47] D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018.

[48] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of Computational Statistics*, 2010, pp. 177–186.

[49] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.

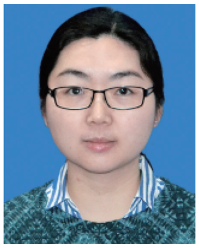
[50] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[51] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[52] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[53] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[54] W. Jin, J. Xu, M. Cheng, Y. Zhang, and W. Guo, "Icnet: Intra-saliency correlation network for co-saliency detection," *Advances in Neural Information Processing Systems*, 2020.



Zhengzheng Tu received the M.S. and Ph.D.degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision and deep learning.

Zhun Li received the B.Eng. degree in Anhui University, in 2019. He is pursuing M.S. degree at the School of Computer Science and Technology, in Anhui University, Hefei, China. His current research interests include computer vision and deep learning.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral research Fellow with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. He was the recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning and deep learning.